DOCUMENT RESUME

ED 406 187 SE 059 877

AUTHOR Boone, William J.; And Others
TITLE Race, Gender, Test Length, and Missing Data: Why Estimates of Performance May Be Clouded.
PUB DATE 22 Mar 97
NOTE 9p.
AVAILABLE FROM Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (Chicago, IL, March 22, 1997).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Data Analysis; Educational Change; Intermediate Grades; Junior High Schools; Middle Schools; *Minority Groups; *Racial Differences; *Sex Differences; *Test Results
IDENTIFIERS Middle School Students

ABSTRACT

In 1991, Ohio received National Science Foundation (NSF) funding through its Statewide Systemic Initiative (SSI) program. One aspect of the reform effort involved evaluating the performance of middle school students with a test item bank of items from the National Assessment of Educational Progress (NAEP). This paper presents the results of evaluating these data. It explores how unanswered items can effect analysis of such data when it is used to calculate mean performance measures of groups. How "missing" data can influence calculations of group performance is significant, for if particular subgroups do not complete a test in much higher numbers than other subgroups, it is likely that analyzed data may not reflect reality. Analyzed data showed a great disparity in the percentage of blacks and whites answering the science test items. Noteworthy are black and white students' answering patterns toward the end of the science test. Findings indicate that male and female test takers exhibit some of the same trends as observed in the racial composition. It is concluded that the design of science tests can greatly influence the quality of achievement measures calculated for students. When tests are administered to students, it is critical to evaluate the influence missing data may have upon calculations. (Author/JRH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 406 187

Race, Gender, Test Length, and Missing Data. Why Estimates of Performance may be Clouded.

William J. Boone
Indiana University
School of Education, Room 3130
201 N. Rose Avenue
Bloomington, IN 47405

Steven R. Rogg
Miami University
420 McGuffey Hall
Oxford, OH 45056

Jane Butler Kahle
Miami University
420 McGuffey Hall
Oxford, OH 45056

Arta Damnjanovic
Miami University
420 McGuffey Hall
Oxford, OH 45056

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
W.J. Boone
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presentation at the 1997 Annual National Association for Research in Science Teaching (NARST) Conference in Chicago, March 22, 1997.

BEST COPY AVAILABLE

SE059077

Race, gender, test length, and missing data. Why estimates of performance may be clouded.

William J. Boone, Indiana University

Steve Rogg, Jane Butler Kahle, and Arta Damnjanovic, Miami University

In 1991, Ohio received NSF funding through its SSI program. One aspect of the reform effort involved evaluating the performance of middle school students with a test item bank of NAEP items. This paper presents the results of evaluating these data. Specifically, how unanswered items can/can not effect analysis of such data when it is used to calculate mean performance measures of groups. How "missing" data can influence calculations of group performance (e.g. females -vs- males) is significant for if particular subgroups do not complete a test in much higher numbers than other subgroups it is likely that analyzed data may not reflect reality. If missing data does influence calculation of subgroup science performance, what are the implications with regard to the analysis and the construction of science tests? Analyzed data show a great disparity in the percentage of blacks and whites answering the science tests items. Noteworthy are black and white students' answering (and not answering) patterns toward the end of the science test. At the end of the test the disparity between blacks and whites attempting items increases significantly. Male and female test takers exhibit some of the same trends as observed in the racial comparison.

Race, gender, test length and missing data.

Why estimates of performance may be clouded.

Introduction:

In 1991, Ohio was one of the first ten states to receive National Science Foundation funding through its new Statewide Systemic Initiative (SSI) program. Because of Ohio's size and large population, its effort was deliberately restricted to middle school (grades five through nine) science and mathematics. Further, it focused on practicing teachers for whom it provided sustained professional development. Four years into the reform, a study was implemented to describe progress, particularly to assess administrator, teacher, and parent attitudes, teaching practices, and student learning. It attempted to describe the landscape of science and mathematics education in Ohio and, hence, was called the Landscape Study (Kahle and Rogg, 1995). This paper focuses on one component of the collected student data.

Objective:

One aspect of the reform effort carried out during Ohio's SSI was to evaluate the performance of students using a set of well-piloted NAEP items. The objective of this paper is to present the results of evaluating the science test item data. Specifically, how unanswered items can/can not effect analysis of such data when it is used to calculate mean performance measures of groups as a function of race and gender. The issue of how "missing" data can influence calculations of group performance (e.g. females -vs- males) is significant in science education for if one particular subgroup does not complete a test in much higher numbers than

other subgroups it is likely that the picture painted with analyzed data may not reflect reality. If missing data does influence calculation of subgroup science performance, then what are the implications with regard to the analysis and the construction of science tests?

Design & Analysis:

In 1996 a 28 item science test was administered to a random sample of 1866 students throughout the state of Ohio. The sample consisted of 520 african-americans and 1346 whites. The breakdown in terms of gender was 1008 females, and 858 males. Following data collection the responses were evaluated utilizing a probabilistic model (Rasch, 1960). This model enabled students performance to be calculated on a linear scale, which allowed parametric tests to be utilized (Wright and Stone, 1979).

Following this analysis, an evaluation of the percentage of students not answering items as a function of race and gender was made. Figure 1 presents the results of the evaluation as a function of gender, while figure 2 presents the data as a function of race.

Findings & Significance:

The data presented in figure 1 are significant for they show that there is a great disparity in the percentage of blacks and whites answering the science tests items. Perhaps most noteworthy are black and white students' answering (and not answering) patterns toward the end of the science test. At the end of the test the disparity between blacks and whites attempting items increases significantly. The data comparing the answering of items by male and female test takers is presented in figure 2. The gender data exhibits some of the same trends as observed

in figure 1. However, there are some differences as well. Figure 2 shows that during the early and mid parts of the test there is no significant difference in the percentage of males and females answering/not answering the test items. The difference in test item answering pattern is apparent however, toward the end of the test. Once the answering patterns past item 22 are examined, it becomes apparent that a significantly greater percentage of females do not answer items than the males.

Implications:

There are important implications of the patterns present in figure 1 and 2. Implication #1: When the final items in tests similar in construction and length to this test are very difficult, then the counting of unanswered items as wrong will not greatly effect the overall performance calculated for a group that does not complete the final items of a test. In this case the overall performance of the females would not have been poorly estimated when a comparison was made to males. Implication #2: When the items at the end of a test of similar length and construction are "easy" then the counting of "not answered" items against test takers could greatly effect the performance measure calculated for males and females. In the case in which "easy" items at the end of a test are not answered by females, the net effect is that their performance is underestimated. Clearly, the same situation exists when the performance of african americans is compared to that of whites.

Conclusion:

The design of science tests can greatly influence the quality of achievement measures calculated for students. When tests are administered to students, it is critical to evaluate the influence missing data may have upon calculations.

In many cases the counting of "not attempted items" as wrong on science tests can cause the following problems:

1. Underestimate performance of slower test takers.
2. Overestimate the achievement gap between females-males, whites-african americans.

Although these pitfalls could have effected Ohio's SSI data a key step way taken - statistical tests that did not count missing data as wrong were used. This meant that the test taking pattern of respondents would not influence the achievement measures calculated for each individual student. For others calculating "science achievement", similar methods should be utilized unless test taking strategies are also being evaluated. Also a range of science item difficulties should be present throughout tests. This would mean that neither a preponderance of only easy or only hard items would not be attempted by students, solely because of location within a test.

A second key issue not only has to do with the placement of items throughout a test as a function of difficulty, but also with regard to the distribution of items which might define one of many subscales on a test. If science test items are being used not only for an overall "science" measure, as well as subscale measures - then the items defining the subscale should be evenly placed throughout the test. If this is not done, then one subscale may have items that were not attempted in greater numbers than other subscales.

References:

Kahle, J.B. and Rogg, S.R. 1995. A Pocket Panorama of the Landscape Study. Oxford, OH: Miami University.

Rasch, G. 1960. Probabilistic Models for Some Intelligence Tests. Danmarks paedogogiske Institute. (Reprinted, University of Chicago Press, Chicago, 1980)

Wright, B.D. and Stone, M.H. 1979. Best Test Design. University of Chicago, MESA Press.

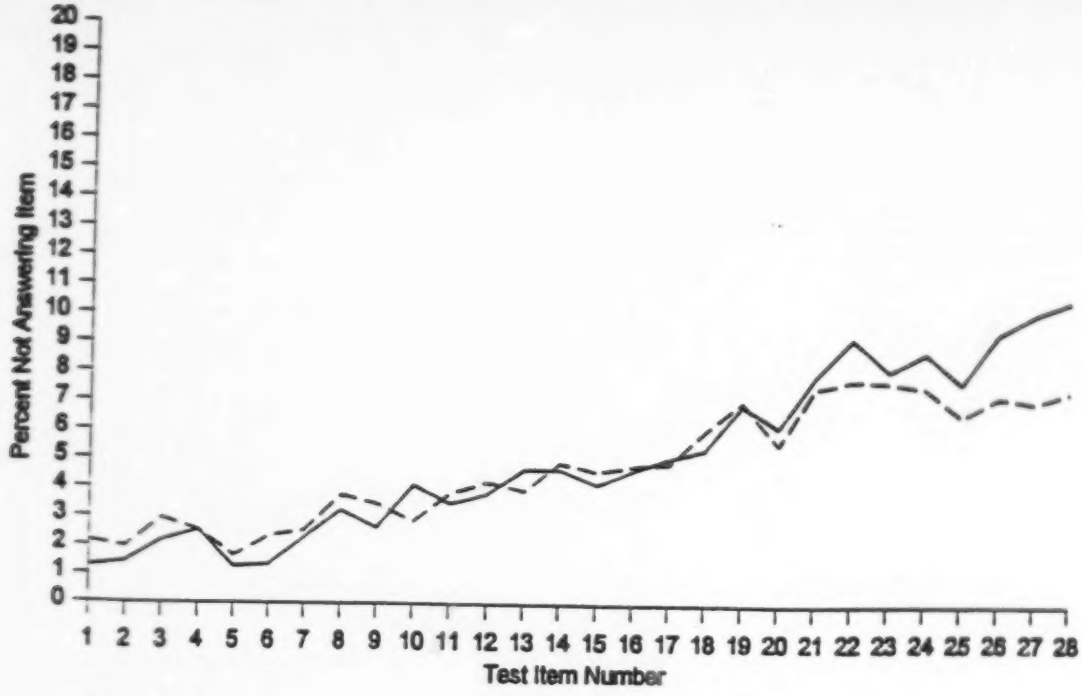
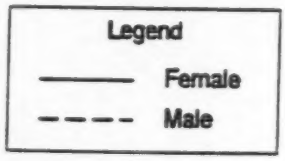


Figure 1



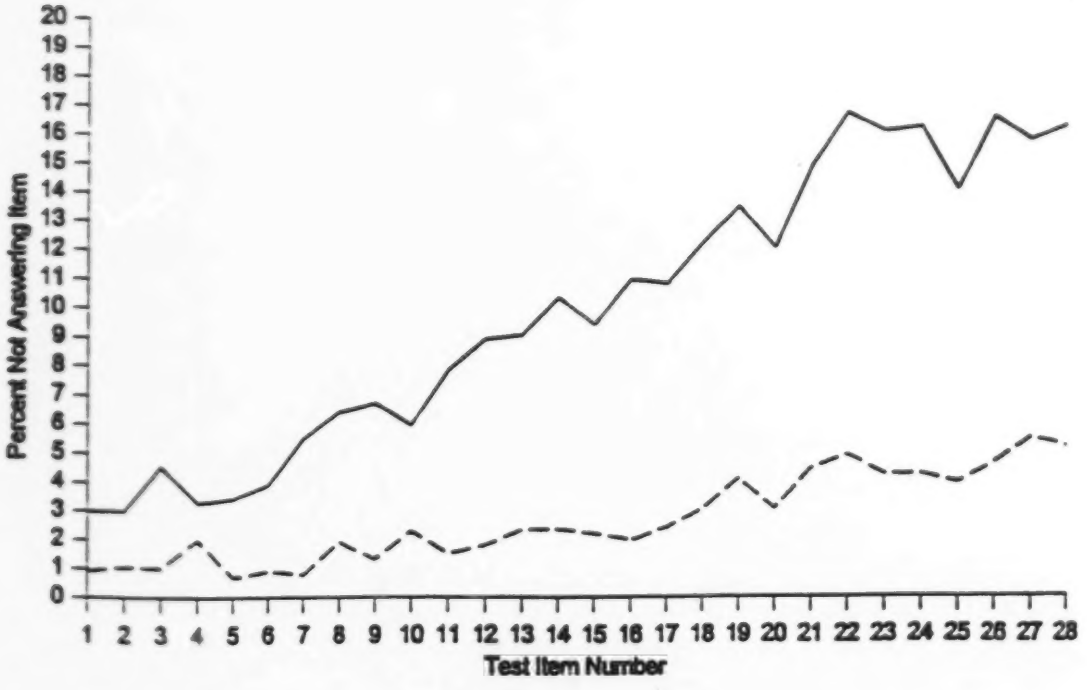


Figure 2